National Statistical Service

Confidentiality Information Sheet 1

What is confidentiality?
An obligation to the provider of information to maintain the secrecy of that information.
Source: UN Economic Commission for Europe, 2009

Confidentiality

What is it and why is it important?

Agencies collecting information from people and organisations have a legal and ethical responsibility to ensure:

- they respect the privacy of those providing the information; and
- that individuals and organisations cannot be identified in a disseminated dataset.

There is a clear relationship between confidentiality and privacy. A breach of confidentiality can result in disclosure of information which might intrude on the privacy of a person or an organisation.

Confidentiality refers to the obligation of data custodians (agencies that collect information) to keep the confidential information they are entrusted with secret.

Why is confidentiality important?

Agencies collecting data often rely on the trust and goodwill of the Australian people to provide information.

Maintaining public trust helps to achieve better quality data and a higher response to data collections.

Protecting confidentiality is a key element in maintaining the trust of data providers.

This leads to reliable data to inform governments, researchers and the community.

Confidentiality and therefore trust can be broken when a person or organisation can be identified in a disseminated dataset, either directly or indirectly.

For example, a person could be **directly** identified in a dataset if that dataset contains their name and address. However, a person or an organisation could also be **indirectly** identified if there is a combination of information in the dataset from which their identity can be deduced.

Example: the combination of date of birth and a detailed area code (for example, a town where 300 people live) may enable identification as there will be some unique dates of birth in such a small area.

What does 'confidentialise' mean?

The term **confidentialise** refers to the steps a data custodian must take to mitigate the risk that a particular person or organisation could be identified in a dataset, either directly or indirectly. Confidentialisation requires two key steps:

1. de-identification of the data, that is, the removal of any direct identifiers (e.g. name and address) from the data; and
2. assessment and management of the risk of indirect identification occurring in the de-identified dataset.

De-identified data does not necessarily protect the identity of individuals or organisations.

Removing identifying information such as name and address protects data providers from **direct identification**.

However, it may still be possible to **indirectly identify** a person or an organisation in a de-identified dataset. If enough detail is available, the identity of a particular person or organisation may be derived from the presence of a very rare characteristic or the combination of unique or remarkable characteristics.

Example: the identity of a person could be deduced if a dataset indicates the person is over 85 years old, has yearly income of more than one million dollars, and resides in a town of 400 people.

Confidentiality – what is it and why is it important

Example: the identity of a person with a very rare disease or health condition could be deduced even in highly aggregated data.

To protect the identity of individuals and organisations, both direct and indirect identification need to be considered.

Confidentialising data involves removing or altering information, or collapsing detail, to ensure that no person or organisation is likely to be identified in the data (either directly or indirectly).

There are various methods used to confidentialise data. These methods aim to protect the identity of individuals and organisations while enabling sufficiently detailed information to be released to make the data useful for statistical and research purposes.

The main techniques for confidentialising data are described in *Confidentiality Information Sheet 4: 'How to confidentialise data: the basic principles'.*

For more information about assessing and managing the risks of indirect identification in microdata see *Confidentiality Information Sheet 5: 'Managing the risk of disclosure in the release of microdata'.*

The confidentiality information series

This information sheet is part of a series designed to explain, and provide advice on, a range of issues around confidentialising data, comprising:

- ***Sheet 1****: 'Confidentiality: what is it and why is it important?';*
- ***Sheet 2****: 'Confidentiality: the obligation to protect identity and privacy';*
- ***Sheet 3****: 'Confidentiality: managing identification risks';*
- ***Sheet 4****: 'How to confidentialise data: the basic principles';*
- ***Sheet 5****: 'Managing the risk of disclosure in the release of microdata';* and
- ***Glossary***.

This series will be expanded in the future to provide further information about aspects of confidentiality.

For more information about confidentiality, or to provide feedback on this series, please email: **inquiries@nss.gov.au**

Definitions

Anonymised data is most commonly used to refer to data from which direct identifiers have been removed, but is sometimes also used to refer to confidentialised data. To avoid confusion, the more specific terms 'de-identified data' and 'confidentialised data' are used in this information series.

Confidentialise — to remove or alter information, or collapse detail, to ensure that no person or organisation is likely to be identified in the data (either directly or indirectly).

De-identified data are data that have had any identifiers removed. De-identified data may also be referred to as **unidentified** data.

Direct identification occurs when identifiers are included with the data that can be used to establish the identity of a person or an organisation.

An identifier (also referred to as a direct identifier) is information that directly establishes the identity of a person or an organisation. For example, name, address, driver's licence number, Medicare number or Australian Business Number.

Indirect identification occurs when the identity of a person or an organisation is disclosed, not through the use of direct identifiers, but through a combination of unique characteristics.

Personal information is *'... information or an opinion (including information or an opinion forming part of a database), whether true or not, and whether recorded in a material form or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion.' (Privacy Act, 1988)*

For more definitions see *Confidentiality Information Sheet – 'Glossary'.*